IMPROVED PARALLEL MODEL COMBINATION TECHNIQUES WITH SPLIT GAUSSIAN MIXTURES FOR SPEECH RECOGNITION UNDER NOISY CONDITIONS

Jeih-weih Hung[1,2], Jia-lin Shen[1] and Lin-shan Lee[1,2]
[1]Institute of Information Science, Academia Sinica
[2]Dept of Electrical Engineering, National Taiwan University
Taipei, Taiwan, Republic of China
Email : {jwhung, jlshen}@iis.sinica.edu.tw

ABSTRACT

The parallel model combination (PMC) technique has been very successful and frequently used to improve the performance of a speech recognition system under noisy environments. In this approach it is assumed that the log spectrum of speech signals is Gaussian-distributed, which is not always valid especially when the number of mixtures in the HMM's is few. In this paper, a simple approach is proposed to improve the PMC method by splitting the mixtures before the domain transformation process in PMC is performed, and merging the mixtures back to the original number after the PMC processes are completed. Preliminary experimental results show that the increased number of mixtures during the PMC processes can in fact provide significant improvements over the original PMC method in terms of the recognition accuracies, especially when the SNR is low.

1. INTRODUCTION

The performance of a speech recognition system often degrades seriously in a noisy environment due to the mismatch between the training and recognition conditions. In order to compensate for the mismatch and achieve better recognition results in different noise conditions, various approaches have been proposed. They can be roughly divided into two categories: compensation with respect to (1) the features and (2) the models. Parallel Model Combination (PMC) has been a very successful technique of the latter category for the environment of additive noise. In this approach the approximated HMM's under noisy conditions can be derived if the HMM's for the additive noise and the clean speech are available. This technique makes possible the avoidance of collecting enormous noisy speech data in order to retrain the HMM's matched to each noisy condition, because only a small amount of noise data are required to train the noise HMM for each noisy condition.

The PMC method has been shown to provide significantly improved recognition accuracy under various noisy environments. However, it is also found that the approximated noisy speech HMM's obtained by PMC still can't work as well as the matched noisy speech HMM's directly trained by the noisy speech, and the performance gap between them becomes wider especially when the SNR becomes worse. There may be many sources for the performance degradation. One example is the inaccuracy occurred in the combination process of PMC, when a correlation term between speech and noise was assumed to be zero [2]. Another example is the inaccuracy occurred when the HMM's cepstral parameters are transformed into the linear spectral domain in PMC [3]. In the domain transformation process the assumption that the probability distribution of speech signal is Gaussian in the log-spectral or cepstral domain has been shown to be not always valid [1]. Some modifications of the PMC techniques have also been developed along this direction [3]. In this paper, we proposed a new simple approach to improve the PMC method by splitting the log-spectral mixtures before they are transformed into the linear spectral domain, so that more accurate noisy speech HMM's can be obtained and the recognition accuracy can be improved significantly.

The remainder of the paper is organized into 3 sections. In section 2, the transformation problem of the PMC method mentioned above will be discussed, and its modified approach will also be proposed. Section 3 then presents some experimental results using this proposed approach. Finally, a short conclusion is given in section 4.

2. THE LIMITATIONS OF PMC AND THE PROPOSED APPROACH

The main concept of the current PMC technique is that since the noise is additive to the speech in the linear spectral domain, the cepstral-based parameters of the clean speech HMM's and the noise HMM must be transformed into the log-spectral domain or the linear spectral domain in order to perform the combination. The parameters for the combined models are then inversely then transformed back to the cepstral domain for the normal recognition process. If the noise HMM is one state and one mixture per state, the number of states and mixtures of the resulting PMC-derived noisy speech HMM's will be the same as those of the original clean speech HMM's. In such cases the recognition process complexity will not be increased at all. Also, the observation within a state of a HMM is often modeled as Gaussian-distributed. Although Gaussian is not always a good model for the speech signal characteristics, multiple mixtures of Gaussian components can describe the speech signal more accurately. It's also widely known that within a reasonable extent more mixtures per state gives higher recognition accuracy.

In the PMC method, the cepstral domain clean speech HMM's must first be transformed into the log-spectral domain, and then into the linear spectral domain, where the latter domain transformation is performed based on the assumption that the logarithm of speech signals is Gaussian-distributed. That is, if the log-spectrum x^l of the speech signal x is Gaussian-distributed as $N\left(\mu_{x^l}, \sigma^2_{x^l}\right)$, then the mean value of the speech signals in linear-spectral domain can be obtained as:

$$\mu_x = E\left(e^{x^l}\right) = \frac{1}{\sqrt{2\pi\sigma_{x^l}^2}} \int_{-\infty}^{\infty} e^{x^l} \exp\left(-\frac{\left(x^l - \mu_{x^l}\right)^2}{2\sigma_{x^l}^2}\right) dx^l$$
$$= \exp\left(\mu_{x^l} + 0.5\sigma_{x^l}^2\right) \left[\frac{1}{\sqrt{2\pi\sigma_{x^l}^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{\left(x^l - \mu_{x^l} - \sigma_{x^l}^2\right)^2}{2\sigma_{x^l}^2}\right) dx^l\right]$$
$$= \exp\left(\mu_{x^l} + 0.5\sigma_{x^l}^2\right) \tag{1}$$

The covariance of the speech signal can be derived similarly. However, by comparing the values of the parameters for models obtained in this way with those for models trained directly in the linear spectral domain with the same database, it's found that there exists a significant discrepancy between them, and such discrepancy is larger for larger log-spectral variance [3]. In other words, the transformation method used in the current PMC between the log-spectral and linear spectral domains is not very accurate especially when the log-spectral variance is large.

As discussed in the above, larger number of Gaussian mixtures per state in the HMM's not only gives higher recognition accuracy empirically, but can model the signal distribution better with multiple mixtures of Gaussian components in principle. In consequence, less number of mixtures per state may cause the inaccuracy in the domain transformation processes of PMC. Moreover, very often the variance value in each Gaussian mixture decreases with the increase in the number of mixture components, This is also a good reason why the discrepancy between the PMC-derived model and the model trained by matched noisy speech data becomes larger for larger values of the log-spectral variance.

With the above observations and considerations, a new simple approach is developed in this paper based on splitting the Gaussian mixtures. This approach will improve the effectiveness and accuracy of the current PMC method without changing its original derived formula. The basic idea is stated as follows. Before a log-spectral HMM mixture is transformed into the linear domain, it can first be "split" into two mixtures by some simple algorithm. Following the original PMC method, the split two mixtures are individually transformed into linear spectral domain, combining with the linear-spectral noise HMM, and transformed back into the log-spectral domain. These two log-spectral noisy mixtures are then combined into one mixture, and finally transformed back into the cepstral domain for normal recognition process. Because the number of mixture components is doubled, the Gaussian distribution assumption can model the speech signal better, and therefore the domain transformation of PMC method can be more accurate. Also, because the two split mixtures are finally combined into one, the total number of mixtures involved in the recognition processes and thus the computation complexity remain unchanged.

A simple algorithm to split a mixture into two is as follows. For a distribution $f(\mathrm{x})$ with mean and covariance $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively, if it is composed of two Gaussian multivariate mixtures, $N(\boldsymbol{\mu}_1, \Sigma_1)$ and $N(\boldsymbol{\mu}_2, \Sigma_2)$ with the same probability,

$$f(\mathbf{x}) = 0.5N(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.5N(\mathbf{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \tag{2}$$

$$\boldsymbol{\mu} = 0.5\boldsymbol{\mu}_1 + 0.5\boldsymbol{\mu}_2, \tag{3}$$

and

$$\boldsymbol{\Sigma} = 0.5(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) + 0.5(\boldsymbol{\mu} - \boldsymbol{\mu}_1)(\boldsymbol{\mu} - \boldsymbol{\mu}_1)^T + 0.5(\boldsymbol{\mu} - \boldsymbol{\mu}_2)(\boldsymbol{\mu} - \boldsymbol{\mu}_2)^T \tag{4}$$

If we assume that

$$(\boldsymbol{\mu}_1)_i = (\boldsymbol{\mu})_i + \alpha(\boldsymbol{\Sigma})_{ii}, \quad (\boldsymbol{\mu}_2)_i = (\boldsymbol{\mu})_i - \alpha(\boldsymbol{\Sigma})_{ii}, \tag{5}$$

where $(\boldsymbol{\mu}_j)_i$ and $(\boldsymbol{\mu})_i$ are the *i*-th component of $\boldsymbol{\mu}_j$ and $\boldsymbol{\mu}$ respectively, *j*=1, 2, $(\boldsymbol{\Sigma})_{ii}$ is the (*i*, *i*) element of $\boldsymbol{\Sigma}$, and the splitting parameter α determines the degree of split, $0 \le \alpha < 1$, and let $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ for simplicity, then

$$(\boldsymbol{\Sigma}_1)_{ij} = (\boldsymbol{\Sigma}_2)_{ij} = \boldsymbol{\Sigma}_{ij} - \alpha^2 \sqrt{\boldsymbol{\Sigma}_{ii}\boldsymbol{\Sigma}_{jj}} \tag{6}$$

Therefore, equations (5) and (6) can be used to split a log-spectral mixture into two. After the two mixtures are updated by the PMC method, equations (3) and (4) can be used to combine them into one. Figure 1 illustrates the complete process of the proposed method.

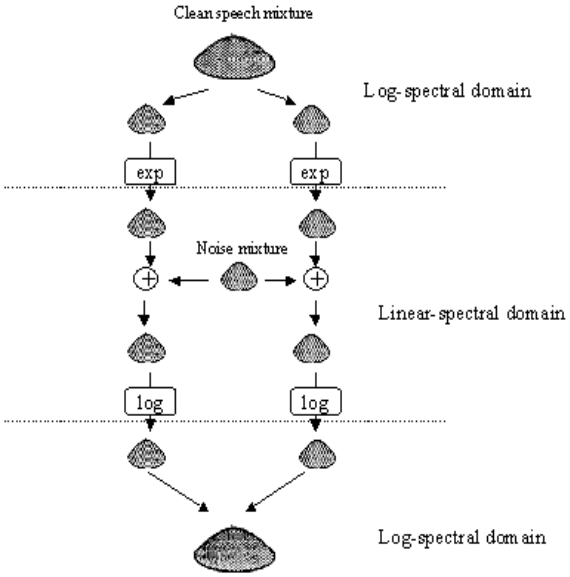

Figure 1. The process of the mixture splitting method

3. EXPERIMENTAL RESULTS

Some initial computer simulation experiments were performed to verify the proposed approach mentioned here. The training speech database used in the experiments contains 3 sets of the 1345 isolated syllables in Mandarin Chinese for each speaker. It was used to train 113 right context-dependent (RCD) INITIAL HMM's and 41 context-independent (CI) FINAL HMM's. Another set of 1345 syllables for the same speaker was used as the test data to be recognized in speaker dependent mode. The results below are the average of 3 speakers. 14 order mel-frequency cepstral coefficients were used as the feature parameters. Also, the clean speech continuous density HMM (CHMM) is trained with 1 state per INITIAL model and 2 states per FINAL model. Two versions of clean speech CHMM were with 2 mixtures per state and 1 mixture per state respectively. Noise HMM's are also individually trained for different levels of white noise and F16 noise to be added to the clean speech, composed of one state and one mixture per state.

First, based on the clean speech HMM's with 1 mixture per state, Table 1 shows the recognition accuracies using the clean speech HMM's, the original PMC and the proposed method with different values of splitting parameter α. It can be found from this table that the clean speech HMM's always give relatively poor performance, the original PMC provides better accuracy, while with the proposed method, the recognition rates are always significantly improved compared with the original PMC, and as the value of the splitting parameter α increases, the improvements become even more significant. Such a trend can be observed for both white noise and F16 noise under all different SNR conditions.

		White noise			F16 noise		
SNR		10dB	20dB	30dB	10dB	20dB	30dB
Original PMC (α=0)		11.97	25.65	55.84	9.29	41.04	70.30
Mixture splitting before PMC	α=0.3	15.39	32.57	61.49	14.72	45.50	71.08
	α=0.5	18.74	40.82	65.50	17.99	51.90	72.86
	α=0.7	25.87	51.30	70.86	29.59	60.52	76.21
	α=0.8	32.94	56.65	73.46	39.26	65.20	77.92
	α=0.9	38.74	59.18	73.83	48.10	69.22	78.96

Table 1 Recognition accuracies for the clean speech HMM's, the original PMC and the proposed mixture splitting approach with different values of splitting parameter α. The clean speech CHHM's to be updated have 1 mixture per state.

Next, Table 2 shows the recognition results for the same experiments except that the clean speech CHMM's to be updated have 2 mixtures per state. In comparison with the results in Table 1, we see that increasing the number of mixtures per state in a HMM from 1 to 2 apparently improves the recognition rates. Furthermore, exactly the same trend as observed previously in Table 1 can be found in Table 2, except that in Table 2 when α is increased from 0.8 to 0.9, the recognition rates are improved only slightly for F16 noise, and even degraded slightly for white noise.

		White noise			F16 noise		
SNR		10dB	20dB	30dB	10dB	20dB	30dB
Original PMC (α=0)		24.16	44.98	69.81	25.87	58.22	78.74
Mixture splitting before PMC	α=0.3	28.55	49.22	72.42	29.14	60.59	79.26
	α=0.5	33.75	53.75	75.02	35.54	66.17	80.07
	α=0.7	40.59	60.82	76.13	46.77	72.12	82.16
	α=0.8	42.97	63.94	77.62	51.52	75.02	82.09
	α=0.9	42.83	63.57	76.13	53.46	75.09	83.57

Table 2 Recognition accuracies for the clean speech HMM's, the original PMC and the proposed mixture splitting approach with different values of splitting parameter α. The clean speech CHHM's to be updated have 2 mixtures per state.

Finally, Tables 3 and 4 show the recognition rates for 1 and 2 mixtures per state respectively when the mixture splitting approach proposed here is accompanied with the previously proposed modified PMC [3], in which the domain transformation formula of PMC is modified as follows,

$$\mu_x = \exp\left(\mu_{x^l} + 0.5\sigma_{x^l}^2\right)\left[\frac{\sigma_{x^l}}{K}\int_{-\infty}^{a-\sigma_{x^l}} e^{-\frac{y^2}{2}} dy\right] \qquad (7),$$

where μ_{x^l} and $\sigma_{x^l}^2$ are the log-spectral mean and variance, respectively, and μ_x is the transformed linear spectral mean. It can be found that the previously proposed modified PMC performs roughly as well as the currently proposed mixture splitting method. However, the two methods can be combined to give better results. In other words, if the log-spectral mixtures are split before transformed to the linear spectral domain using equation (7), the recognition rates can be further improved.

		White noise			F16 noise		
SNR		10dB	20dB	30dB	10dB	20dB	30dB
Modified PMC (α=0)		32.27	54.72	70.26	42.08	63.87	76.21
Modified PMC with mixture splitting	α=0.3	37.17	57.99	71.75	47.06	68.40	77.99
	α=0.5	37.77	58.44	72.19	49.44	68.77	78.51
	α=0.7	39.41	59.41	73.31	49.37	69.07	78.88
	α=0.8	39.33	59.11	73.31	48.85	70.48	80.22
	α=0.9	38.96	59.63	73.23	47.81	69.44	79.85

Table 3 Recognition accuracies for the modified PMC method combined with the mixture splitting approach with different values of splitting parameter α. The clean speech CHHM's to be updated have 1 mixture per state.

		White noise			F16 noise		
SNR		10dB	20dB	30dB	10dB	20dB	30dB
Modified PMC		43.05	62.53	77.70	48.85	70.93	81.93
Modified PMC with mixture splitting	α=0.3	43.27	63.27	78.36	49.29	72.42	81.71
	α=0.5	43.87	63.72	77.84	50.48	72.79	82.38
	α=0.7	44.46	64.31	78.29	51.30	73.61	82.30
	α=0.8	43.87	64.83	77.77	51.67	73.68	82.68.
	α=0.9	42.16	64.54	77.99	50.63	72.94	82.01

Table 4 Recognition accuracies for the modified PMC method combined with the mixture splitting approach with different values of splitting parameter α. The clean speech CHHM's to be updated have 2 mixtures per state.

4. CONCLUSION

In this paper, it is shown that proper splitting of mixtures before the domain transformation process can produce much better accuracies in the very successful PMC method under noisy conditions, especially when the SNR is low. Furthermore, it is also shown that the recognition performance can be further improved when the proposed mixture splitting technique is combined with the modified PMC previously proposed with better domain transformation processes.

.

REFERENCES

[1] M.J. Gales and S.J. Young, "Robust Speech Recognition in Additive and Convolutional Noise Using Parallel Model Combination", Computer, Speech and Language 9, pp. 289-307, 1995.

[2] Jeih-weih Hung, Jia-lin. Shen and Lin-shan. Lee, "Improved Robustness for Speech Recognition Under Noisy Conditions Using Correlated Parallel Model Combination", 1998 IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '98).

[3] Jeih-weih Hung, Jia-lin Shen, and Lin-shan Lee, "Improved Parallel Model Combination Based on Better Domain Transformation for Speech Recognition Under Noisy Environments", to appear in 1998 Int. Conf. On Spoken Language Processing (ICSLP '98).